CDI15

T2. *Information and entropy*

303 SXD

2.1. Uncertainty, entropy and information

2.2. Properties of the entropy function

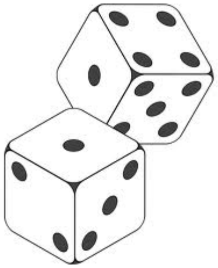
2.3. Conditional entropy

2.4. Readings

2.5. References

2.1. Uncertainty, entropy and information

Uncertainty is present in *random experiments* (or *random variables*), like throwing a coin, or a die, or a pair of dice, or spinning a roulette, or betting about the value of the € against the $ tomorrow, and so on.



Is there a reasonable way to measure such uncertainty?

Notations. $U = \{a_1, a_2, \dots, a_n\}$, the set of possible outcomes of a random (or *stochastic*) variable A. Set $p_j = P(a_j) > 0$, the probability of obtaining a_j. If $X \subseteq U$ (these subsets are called *events*), $p = P(X) = \sum_{a_j \in X} p_j$ is the *probability of* X. Note that $p > 0$ unless $X = \emptyset$ (the *impossible event*). As $P(U) = p_1 + p_2 + \cdots + p_n = 1$, U is the *sure event*.

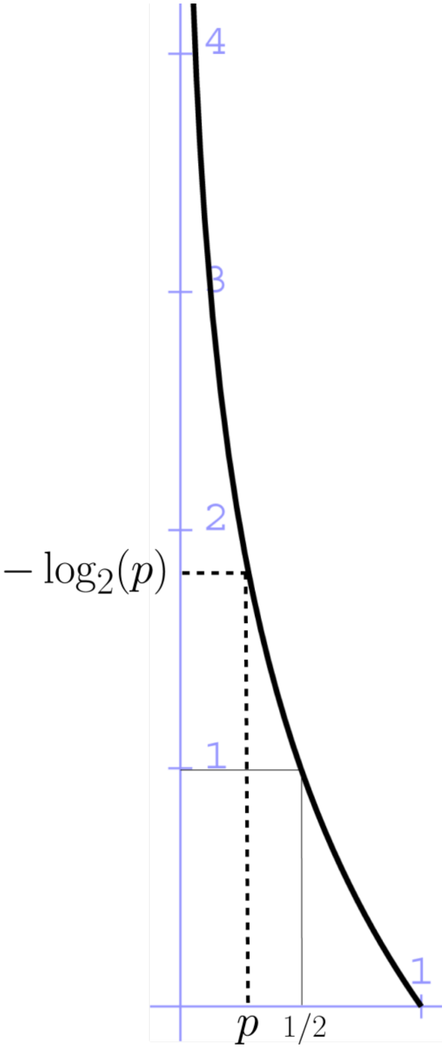
Uncertainty of one event

We look for a measure $H = H(X)$ of the *uncertainty* about the occurrence of an event X, or of the *information* provided by its occurrence.

We will assume that H depends only on p, $H = H(p)$, that $H(p)$ is *continuous* and that:

a) H is *non-negative:*
$H(p) \geq 0$ if $0 < p \leq 1$.

b) H is *additive for independent events*:
$H(pq) = H(p) + H(q)$.



Mathematical magic:

H necessarily has the form

$$H(p) = -k \log p\ ,\ k > 0 \text{ constant.}$$

With the *normalization* $H(1/2) = 1$ (*bit*),

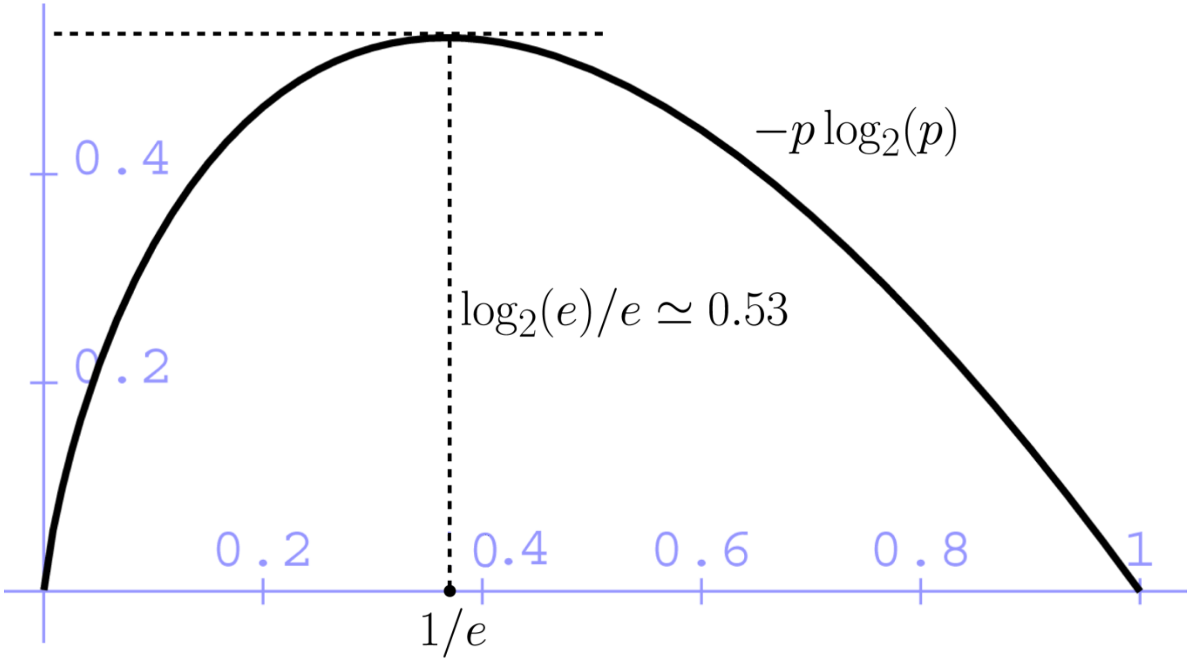
$$H(p) = -\log_2(p).$$

Uncertainty or entropy of A

The *uncertainty* or *entropy* of the random experiment A, $H = H(A)$, is the average uncertainty, or *expected* uncertainty, of its outcomes:

$$H(A) = p_1 H(p_1) + \cdots + p_n H(p_n) = -\sum_j p_j \log_2(p_j).$$

$$H(p_1, \dots, p_n) = -\sum_j p_j \log_2(p_j)$$: Shannon's entropy formula.



The figure shows the graph of the function $-p\log_2(p)$. It has a maximum at $p = 1/e \simeq 0.368$. The value of the maximum is $\log_2(e)/e \simeq 0.53$.

Remark. If the probabilities are given as a list of weights w_i, so that

$$p_k = w_k / w,$$

then we can use the formula

$$H(p_1, \dots, p_n) = \left(-\sum_j w_j \log_2(w_j) + w \log_2(w)\right)/w.$$

It will also be denoted $H(w_1, \dots, w_n)$.

The proof is a simple calculation:

$$\begin{aligned} H(p_1, \dots, p_n) &= -\sum_j p_j \log_2(p_j) \\ &= -\sum_j \frac{w_j}{w} \log_2\left(\frac{w_j}{w}\right) \\ &= -\left(\sum_j w_j(\log_2 w_j - \log_2 w)\right)/w \\ &= \left(-\sum_j w_j \log_2 w_j + \sum_j w_j \log_2 w\right)/w \\ &= \left(-\sum_j w_j \log_2(w_j) + w \log_2(w)\right)/w. \end{aligned}$$

E.2.1. [Welsh-88, p. 3] Which race has greater uncertainty: a handicap in which there are 7 runners, 3 having probability 1/6 and 4 having probability 1/8, or a selling plate in which there are 8 runners with 2 horses having 1/4 probability of winning and 6 horses having each 1/12 probability?

E.2.2. [Welsh-88, p. 11] Which has greater information: 10 letters or 26 decimal digits?

E.2.3. Discuss the saying: *A picture is worth a thousand words.*

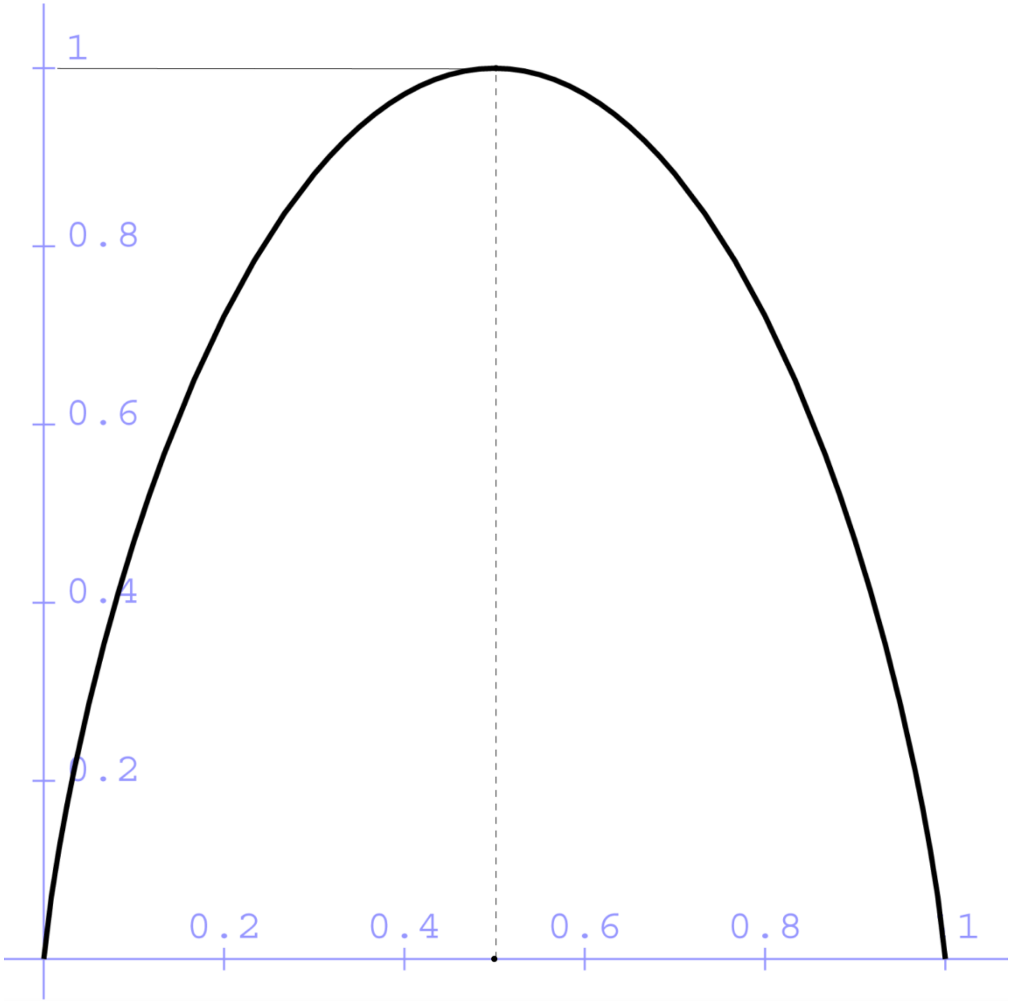
Example. Suppose we have a biased coin with

$$p(\text{heads}) = p, p(\text{tails}) = 1 - p.$$

Then the entropy is

$$c(p) = H(p, 1-p) =$$
$$-p\log_2(p) - (1-p)\log_2(1-p).$$

The figure shows the graph of $c(p)$. Notice that $c(p) > 0$ except for $p = 0$ and $p = 1$, that it is symmetric about $p = 1/2$ and that $c(p)$ has a maximum value, namely 1, at $p = 1/2$.



$c(p) = -p\log_2(p) - (1-p)\log_2(1-p)$

Thus, a fair coin ($p = 1/2$) has the maximum entropy (uncertainty), corresponding to 1 bit of information. This agrees with our intuition that the outcome of throwing a biased coin has less uncertainty than throwing a fair one.

Example. The entropy of an experiment or random variable whose n outcomes are equally likely is $\log_2(n)$.

Indeed, the entropy in question is $H(1/n, 1/n, \ldots, 1/n)$, and by Shannon's formula this is equal to $-\sum_1^n \frac{1}{n}\log_2\left(\frac{1}{n}\right) = n\frac{1}{n}\log_2(n) = \log_2(n)$.

In particular, if we throw a fair coin m times, then the entropy is m bits.

Since $\log_2(n)$ increases when n increases, we see that a random variable with n equally likely outcomes has less uncertainty than a random variable with $m > n$ equally likely outcomes. This agrees with our intuition that the outcome of throwing a *fair* coin is less uncertain than the outcome in throwing a *fair* die, and that this in turn is less uncertain than the result of a (fair) roulette spin.

Remark. Since the uncertainty $H(p)$ is also the information amount gained by the occurrence of an event of probability p, the entropy $H(A)$ can be understood as *the average information gained in a run of* A.

2.2. Properties of the entropy function

a) $H(p_1, p_2, \dots, p_n)$ is a *positive* symmetrical function of $p_1, p_2, \dots, p_n$.
b) $H(p_1, \dots, p_n) \leq \log_2(n)$, and equality is satisfied if an only if

$$p_1 = p_2 = \cdots = p_n = 1/n.$$

This agrees with our intuition that a fair die carries more uncertainty than a biased one. The proof is quite easy using that $\ln(x) \leq x - 1$ *for all* x*, with equality if and only if* $x = 1$ (see the proof on next slide). Indeed, if $q_1, \dots, q_n$ is any other distribution of probability ($q_j > 0, \sum q_j = 1$), then

$\ln(q_j/p_j) \leq q_j/p_j - 1$, with equality if an only if $q_j = p_j$;

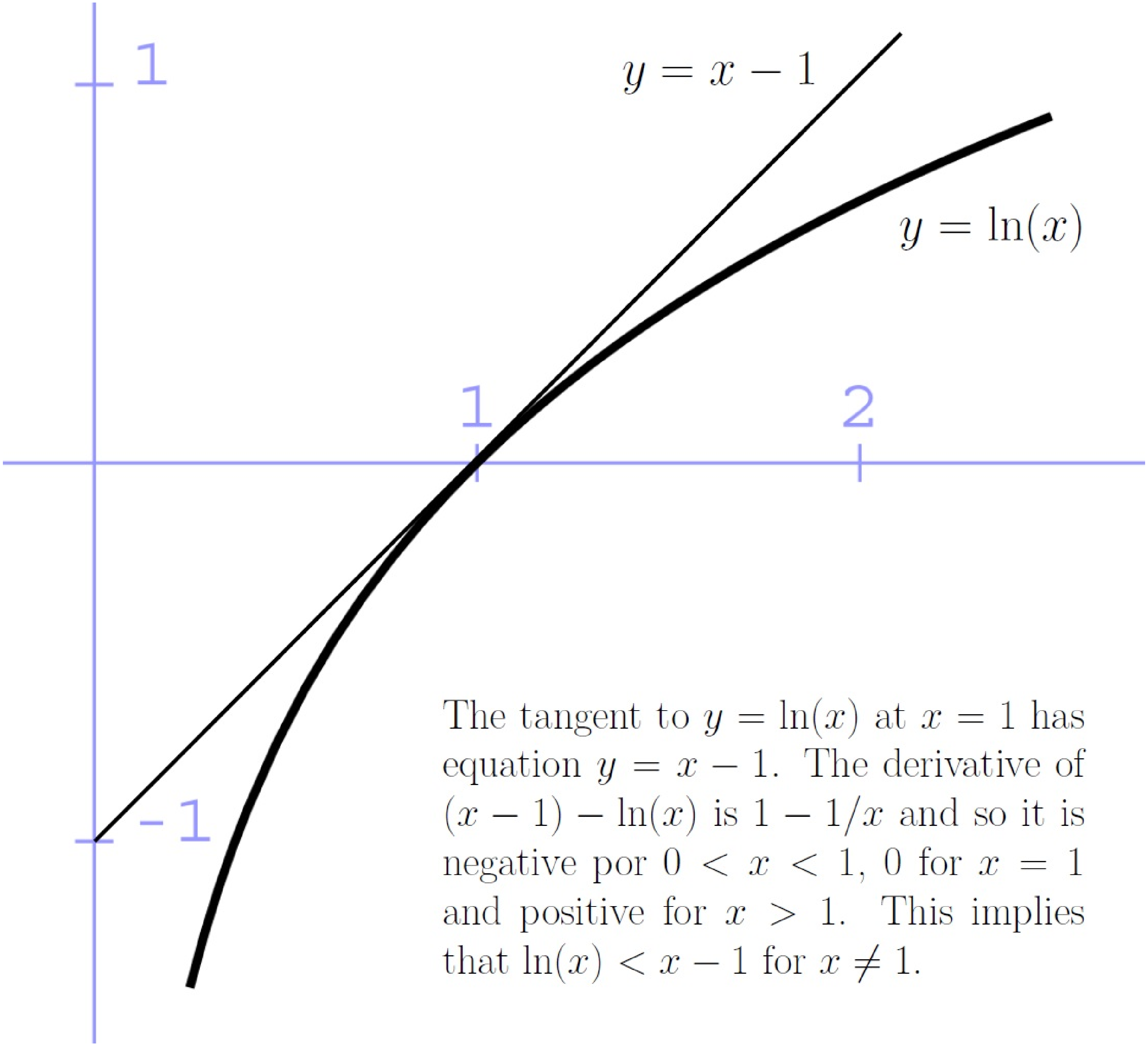
$\sum p_j \ln(q_j/p_j) \leq \sum q_j - \sum p_j = 1 - 1 = 0;$

$-\sum p_j \ln(p_j) \leq -\sum p_j \ln(q_j);$

$H(p_1, \dots, p_n) \leq -\sum p_j \log_2(q_j)$ [multiply line above by $\log_2(e)$];

$H(p_1, \dots, p_n) \leq \log_2(n)$ [set $q_j = 1/n$ in previous line],

with equality if and only if $p_j = q_j = 1/n$ for all j. And that is it!



The tangent to $y = \ln(x)$ at $x = 1$ has equation $y = x - 1$. The derivative of $(x-1) - \ln(x)$ is $1 - 1/x$ and so it is negative por $0 < x < 1$, 0 for $x = 1$ and positive for $x > 1$. This implies that $\ln(x) < x - 1$ for $x \neq 1$.

Remark (Gibbs lemma). In the above proof we have seen that the minimum of $-\sum p_j \log_2(q_j)$ when the $q_1, \dots, q_n$ runs over all possible probability distributions, with the distribution $p_1, \dots, p_n$ fixed, is achieved for $q_j = p_j$. In other words,

$$-\sum p_j \log_2\left(q_j\right) \geq -\sum p_j \log_2\left(p_j\right), \text{ with equality only for } q_j = p_j.$$

c) Let B be another random variable with possible outcomes $V = \{b_1, \dots, b_m\}$ and probabilities $q_1, \dots, q_m$. Then we have the *composite* or *joint* random variable (A, B) whose trials consist in observing A and B together. The possible outcomes are the pairs (a_j, b_k). If we set

$p_{jk} = P\left(a_j, b_k\right) = P\left(A = a_j \wedge B = b_k\right) = p_j P\left(b_k | a_j\right)$, then

$$H(A, B) = -\sum_{jk} p_{jk} \log_2(p_{jk}).$$

There is a relation of this entropy to the entropies $H(A)$ and $H(B)$:

$H(A, B) \leq H(A) + H(B)$ [Upper bound on joint entropy]

with equality if and only if A and B are independent.

	b_1	$\cdots$	b_k	$\cdots$	b_m	
a_1	p_{11}	$\cdots$	p_{1k}	$\cdots$	p_{1m}	p_1
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_j	p_{j1}	$\cdots$	p_{jk}	$\cdots$	p_{jm}	p_j
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_n	p_{n1}	$\cdots$	p_{nk}	$\cdots$	p_{nm}	p_n
	q_1	$\cdots$	q_k	$\cdots$	q_m	1

To prove this, first notice that $\sum_k p_{jk} = p_j$ and $\sum_j p_{jk} = q_k$ (see table). Then

$$\begin{aligned}
&H(A) + H(B)\\
&= -\textstyle\sum_j p_j \log_2(p_j) - \sum_k q_k \log_2(q_k)\\
&= -\textstyle\sum_j \sum_k p_{jk} \log_2(p_j) - \sum_k \sum_j p_{jk} \log_2(q_k)\\
&= -\textstyle\sum_{jk} p_{jk} \log_2(p_j q_k)\\
&\geq -\textstyle\sum_{jk} p_{jk} \log_2(p_{jk}) \ \ \text{[Gibbs lemma, p. 10]}\\
&= H(A,B).
\end{aligned}$$

Equality occurs if and only if $p_j q_k = p_{jk}$, which is precisely the condition for A and B to be independent.

Remark. Given the table of probabilities p_{jk} of a joint distribution, the probabilities $p_j = \sum_k p_{jk}$ and $q_k = \sum_j p_{jk}$ are usually said to be the *marginal distributions* of the joint distribution, by rows and columns, respectively.

Remark. We have, by the formula on conditional probabilities,

$$p_{jk} = P\left(A = a_j \wedge B = b_k\right) = P\left(A = a_j\right)P(B = b_k|A = a_j).$$

In this expression, $P\left(A = a_j\right) = p_j$. As for $P\left(B = b_k \middle| A = a_j\right)$, it is equal to $P(B = b_k) = q_k$ if and only if the event $\{B = b_k\}$ is independent of the event $\{A = a_j\}$. Hence $p_{jk} = p_j q_k$ if and only if $\{B = b_k\}$ is independent of $\{A = a_j\}$. Thus $p_{jk} = p_j q_k$ for all j and k ⇔ A and B are independent.

2.3. Conditional entropy and information

The *conditional entropy* of a random variable B *given an event* X is defined as the uncertainty of the random variable $B|X$:

$$H(B|X) = -\sum_{k=1}^{m} P(b_k|X)\log_2(P(b_k|X)).$$

Remark. $H(B|X) = H(B)$ if and only if B is independent of X.

If A is another random variable, the *conditional entropy* of B *given* A is defined by

$$H(B|A) = \sum_j p_j H(B|a_j).$$

This can be thought as the uncertainty about B that remains after having observed A, averaged with the probabilities p_j of A.

In more detail,

$$H(B|A) = -\sum_j p_j \sum_k P\left(b_k\middle|a_j\right)\log_2\left(P\left(b_k\middle|a_j\right)\right)$$

Remark. Since $H\left(B\middle|a_j\right) = H(B)$ if and only if B is independent of a_j, $H(B|A) = H(B)$ if and only if B is independent of A.

E.2.4. Show that $H(A|A) = 0$.

E.2.5. [Welsh-88, p. 9] For any random variable A show that $H(A^2|A) = 0$ and give an example for which $H(A|A^2) \neq 0$.

Example. Let A and B stand for the input and output bit, respectively, of a binary symmetric channel with cross-over probability p. Then (see the Example on p. 6)

$$H(B|A) = c(p).$$

Indeed, by definition we have

$$H(B|A) = P_A(0)H(B|0) + P_A(1)H(B|1).$$

Now it is enough to notice that

$$P_A(0) = P_A(1) = 1/2,$$

$$H(B|0) = -P_B(0|0)\log_2 P(0|0) - P_B(1|0)\log_2 P(1|0) = c(p),$$

$$H(B|1) = -P_B(0|1)\log_2 P(0|1) - P_B(1|1)\log_2 P(1|1) = c(p).$$

Fundamental formula

$$H(A,B) = H(A) + H(B|A).$$

Proof. $H(A,B) = -\sum_{jk} P\left(a_j, b_k\right) \log_2 \left(P\left(a_j, b_k\right)\right)$

$$= -\sum_{jk} p_j P(b_k|a_j) \log_2\left(p_j P(b_k|a_j)\right)$$

$$= -\sum_{jk} p_j P(b_k|a_j) \log_2\left(p_j\right) - \sum_{jk} p_j P\left(b_k\middle|a_j\right) \log_2 \left(P\left(b_k\middle|a_j\right)\right)$$

$$= H(A) + H(B|A).$$

Here we have used that the log of a product is the sum of the logs and that $\sum_k P(b_k|a_j) = 1.$

Corollary. $H(B|A) \leq H(B)$, with equality if and only if B is independent of A.

Indeed, the fundamental formula and the upper bound on the joint entropy tell us that $H(A) + H(B|A) \leq H(A) + H(B)$, which is equivalent to $H(B|A) \leq H(B)$, with equality holds if and only if B is independent of A.

Conditional information

The *information about B conveyed/provided by* A, $I(B|A)$, is defined by

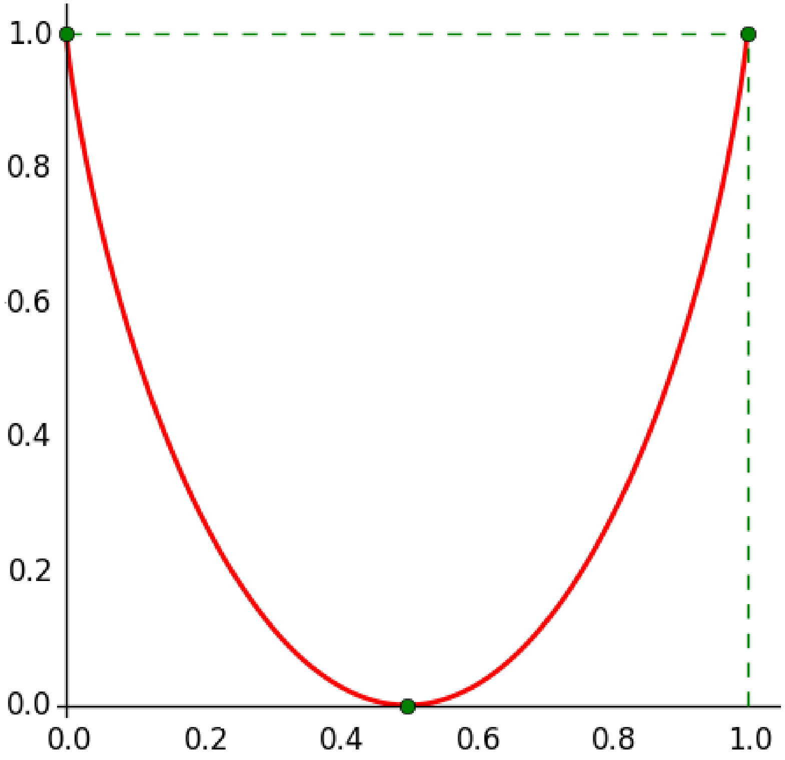
$$I(B|A) = H(B) - H(B|A).$$

Since $H(B|A)$ is the uncertainty about B given A, $I(B|A)$ is the part of $H(B)$ accounted for by observing A. Note that $I(B|A) = 0$ iff B is independent of A. On the other hand, $I(A|A) = H(A)$, as $H(A|A) = 0$ (E.2.1).

Example. Let A and B stand for the input and output bit, respectively, of a binary symmetric channel with cross-over probability p.

Then $I(B|A) = 1 - c(p) = C(p)$ is the *capacity* of the channel (amount of information available at the receiving end for each bit sent). Indeed, $H(B) = 1$ and we have seen (Example on page 14) that

$$H(B|A) = c(p).$$

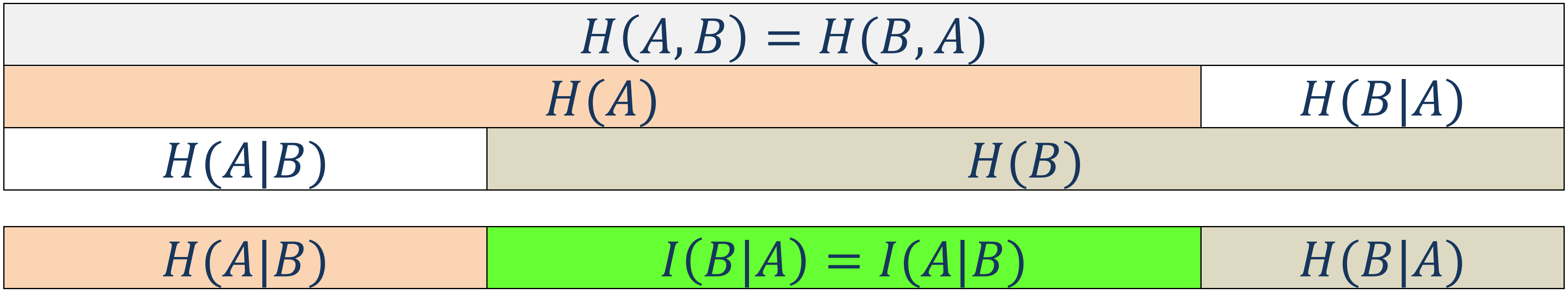


$C(p) = 1 + p\ \log_2(p) + (1-p)\log_2(1-p)$

Corollary. $I(B|A) = I(A|B)$.

Indeed, $I(B|A) = H(B) - H(B|A) = H(B) - (H(A,B) - H(A))$, or

$$\begin{aligned} I(B|A) &= H(A) + H(B) - H(A,B) \\ &= H(B) + H(A) - H(B,A) \\ &= I(A|B). \end{aligned}$$



$H(A,B) = H(B,A)$

$H(A)$ | $H(B|A)$

$H(A|B)$ | $H(B)$

$H(A|B)$ | $I(B|A) = I(A|B)$ | $H(B|A)$

There is another expression about $I(B|A) = I(A|B)$ that is more illuminating about its significance. To this end, recall that given two events X and Y we have the relation (Bayes' rule)

$$P(Y|X) = P(Y) \cdot I(X,Y), \text{ with } I(X,Y) = \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(X|Y)}{P(X)}.$$

If we take $\log_2$, we get

$$-\log_2 P(Y) - (-\log_2 P(Y|X)) = \log_2 I(X,Y)\,. \qquad [*]$$

In this relation, $-\log_2 P(Y) = H(Y)$ is the uncertainty about Y and $-\log_2 P(Y|X) = H(Y|X)$ is the uncertainty about Y assuming we have observed X. Thus the left hand side of $[*]$, namely $H(Y) - H(Y|X)$, is, as we have seen, the information about Y provided by the observation of X. Thus we see that the term $i(X,Y) = \log_2 I(X,Y)$ measures the information about Y supplied by the observation of X (it is positive, negative or zero according to whether $I(X,Y) > 1, < 1$ or $= 1$):

$$H(Y) - H(Y|X) = i(X,Y) \text{ or } H(Y) = H(Y|X) + i(X,Y)\,,$$

which is an additive form of Bayes' rule.

Now we can state and prove the alternative expression for $I(B|A)$:

$$I(B|A) = \sum_{jk} p_{jk} i(a_j, b_k).$$

In other words, the information about B provided by observing A is the weighted average of the informations $i(a_j, b_k)$.

The proof of the formula is a simple computation:

$$\begin{aligned} I(B|A) &= H(B) - H(B|A) \\ &= -\sum_k q_k \log_2 q_k + \sum_j p_j \sum_k P(b_k|a_j) \log_2 \left(P(b_k|a_j)\right) \\ &= -\sum_k \sum_j p_{jk} \log_2 q_k + \sum_j \sum_k p_{jk} \log_2(p_{jk}/p_j) \\ &= \sum_j \sum_k p_{jk} \log_2(p_{jk}/p_j q_k) \\ &= \sum_{jk} p_{jk} i(a_j, b_k). \end{aligned}$$

Note that since $i(a_j, b_k) = i(b_k, a_j)$, this yields another proof of the relation $I(B|A) = I(A|B)$.

E.2.6. Deduce the formula $I(B|A) = 1 - c(p)$ established in the Example on page 15 using the formula on top of this page.

2.4. Readings

Comparing energy and information

The universality and importance of the concept of information could be compared only with that of energy. It is interesting to compare these two (cf. Rényi, 1960a). One is tempted to say that the great inventions of civilization serve either to transform, store and transmit energy (fire, mechanisms like wheels, use of water and wind energy, for instance, for sailing or in mills, steam engines, use of electric, later nuclear energies, rockets, etc.) or they serve to transform, store and transmit information (speech, writing, drum- and fire-signals, printing, telegraph, photograph, telephone, radio, phonograph, film, television, computers, etc.). The analogy goes further. It took a long time (until the middle of the nineteenth century) for the abstract concept of energy to be developed, i.e. for it to be recognized that mechanical energy, heat, chemical energy, electricity, atomic energy, and so on, are different forms of the same substance and that they can be compared, measured with a common measure. What, in fact, remains from the concept of energy, if we disregard its forms of apparition, is its quantity, [its] measure,

> which was introduced some 125 years ago. In connection with the concept of information, this essentially happened a century later, with the works of Shannon (1948a,b). [There is even a "principle of conservation of information"-like that of energy; see Katona and Tusnády (1967) and Csiszar et al. (1969).] Again, if we disregard the different contents (meanings) of information, what remains is its quantity, [its] measure.
>
> [Aczel-Daroczy-75], p. 1-2.

For the relation between Shannon's formula and the thermodynamic entropy, see the Appendix.

2.5. References

[Aczel-Daroczy-1975] J. Aczél, 2. Daróczy: *On measures of information and their characterizations.* Academic Press, 1975.

[Welsh-1988] Dominic Welsh: *Codes and cryptography.* Oxford Sci.Publ., 1988.

[Poynton-2003] Charles Poynton: *Digital Video and HDTV. Algorithms and Interfaces*. Morgan Kaufmann Publishers, An imprint of Elsevier. xlii+794p.

[Sayood-2006] Khalid Sayood: *Introduction to data compression* (3rd edition). Morgan Kaufmann, An imprint of Elsevier, 2006. xxii+680p.

Appendix (adapted from Section 11.2 of S. Haykin's *Neural Networks and Learning Machines* (3rd edition), Pearson, 2009).

Let $\{i\}$ be a set of indices for the internal states of a system composed of a large number of particles. Let E_i denote the energy of the system in the state i. If we can assume that the system is in thermal equilibrium with its surrounding environment, then the probability that the system is in state i is given by the so-called *Gibbs distribution*:

$$p_i = \frac{1}{Z} e^{-E_i/kT}$$

where Z is a constant independent of all states, T is the absolute temperature (in kelvins) and $k = k_{\mathrm{B}} = 1.38 \times 10^{-23} \text{joules/kelvin}$ is *Boltzmann's constant*. Note that $\sum p_i = 1$ implies that

$Z = \sum_i e^{-E_i/kT}$ (this expression is called the *partition function*).

From the expression giving p_i and the properties of the exponential function, we see that p_i increases when E_i decreases, so low energy states are more probable than high energy states (for a given T).

Note also that

$$\log p_i + \log Z = -E_i/kT \text{ or}$$

$$kT\log p_i + kT\log Z = -E_i.$$

If we multiply the last relation by p_i and sum over i, we get

$$-TS + kT\log Z = -\langle E\rangle,$$

where $S = -k\sum_i p_i \log p_i$ (*Gibbs entropy*) and $\langle E\rangle = \sum_i p_i E_i$ (*average energy*). Thus we can conclude that

$$\langle E\rangle = TS + F, \text{ or } F = \langle E\rangle - TS, \qquad [*]$$

where $F = -kT\log Z$ is the so called *free energy* (Helmholtz). But [$*$] is the classical thermodynamical relation defining entropy (S), so we can conclude that the classical entropy is equal to the Gibbs entropy, which itself is proportional to Shannon's entropy. This explains the deep reason for why Shannon chose the word entropy to name the quantity $\sum_i p_i \log p_i$.